Exploring the Geography of Communities in Social Networks

Alexis Comber[1], Michael Batty[2], Chris Brunsdon[3], Andrew Hudson-Smith[2], Fabian Neuhaus[2] and Steven Gray[2]

[1]Department of Geography, University of Leicester, Leicester, LE1 7RH UK
Tel. +44 116 252 3812/ 3823 Fax +44 116 252 3854
ajc36@le.ac.uk, http://www2.le.ac.uk/departments/geography/people/ajc36
[2] Centre for Advanced Spatial Analysis, University College London, W1T 4TJ UK
mbatty@geog.ucl.ac.uk, a.hudson-smith@ucl.ac.uk, fabian.neuhaus@ucl.ac.uk, steven.gray@ucl.ac.uk
[3]Department of Geography, University of Liverpool, Liverpool, L69 3BX UK
christopher.brunsdon@liverpool.ac.uk, http://www.liv.ac.uk/geography/staff/brunsdon.htm

Summary: This research analyses social network data to identify communities or sub-graph regions. These sub-graph areas are indentified based on the arrangement of edges between vertices. The geographies of the communities are analysed, compared and visualised using kernel density estimations. A research agenda is suggested.

KEYWORDS: geo-code, sub-graph, geography, statistical analysis

1. Introduction

This paper develops a spatial analysis of social network data. A vast body of research analysing structure in networks and graphs exists in order to identify 'communities' or sub-graph areas that homogenous in some respect. One of the recurring themes in the community detection research concerns the reliability of the communities that are identified and the difficulty in understanding what they mean. For example, Porter (2009, p1098) notes that "few methods have been developed to use or even validate the communities that we find" and Newman (2008) states that "methods for understanding what the communities mean after you find them are ... still quite primitive" (Newman, 2008, p38). Much social network data now has a geographical component and it is possible to explore how the geographies different communities and partitioning methods compare. Given these contexts, the objectives of this research were:

i) To apply community detection algorithms to social network data;
ii) To analyse the how virtual communities defined in social network space and social network structures relate to geographic space;

2. Case studies

Real networks tend to have a wide distribution of the number of edges connected to each vertex. Thus they are highly heterogeneous with specific sub-graph regions having high concentrations of interconnected vertices. The basis of community detection is to identify areas of the network with higher than expected concentrations of edges connecting groups of vertices and with lower than expected concentrations of edges between these groups. These areas can be considered as 'communities' (Girvan and Newman, 2002). An important area of research in network sciences has been the development of methods for partitioning networks to identify communities. One of the primary aims of these activities is to identify areas within the graph (sub-graph areas) where the nature of interactions between vertices indicates some local clustering of interactions. The assumption is that sub-graph areas with high internal interactions are homogenous to some degree. The interested reader is directed to number of reviews of the methods arising from statistical physics (Porter et al.,

2009; Fortunato, 2010; Leicht & Newman, 2008)). The case studies consider the identification of sub-graph areas from 4 different perspectives: the connectedness of concepts, communities based on interaction, user geography and interaction content. In the case studies specific tags in communications between social network data users ('@user' in Twitter data in this case) are used to identify and illustrate the connectedness of different concepts. The network is defined by the interactions (edges) between users (vertices). A subset of the data was analysed. It contained 87,555 records. Of these 52,397 contained tweets at ('@') a specific user, 52,280 that were not self tweets and 11,968 at users with a spatial reference (Geotag). At the end of the data cleaning, the network comprised 6,659 vertices and 7,491 edges (Figure 1).

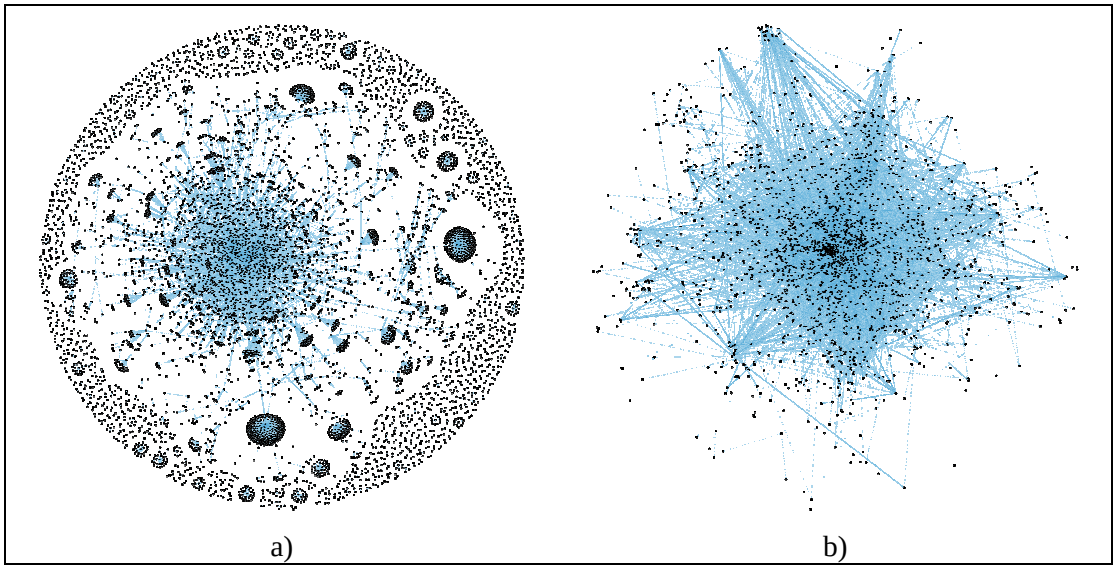


Figure 1. a) The network of users, displayed using a Fruchterman-Reingold layout b) The network displayed over geographic space.

3. Results

The Walktrap or Random Walk algorithm (Pons and Latapy, 2005) was used to partitioned the network into sub-graph areas based. It identified 1181 communities and in total, 33 communities contain more than 20 vertices. These communities can be visualised in geographic and non-geographic space, and the users / vertices that perform 'hub' functions identified by their degree value – the 'degree' of each vertex is the number of edges connected to it (Figure 2).

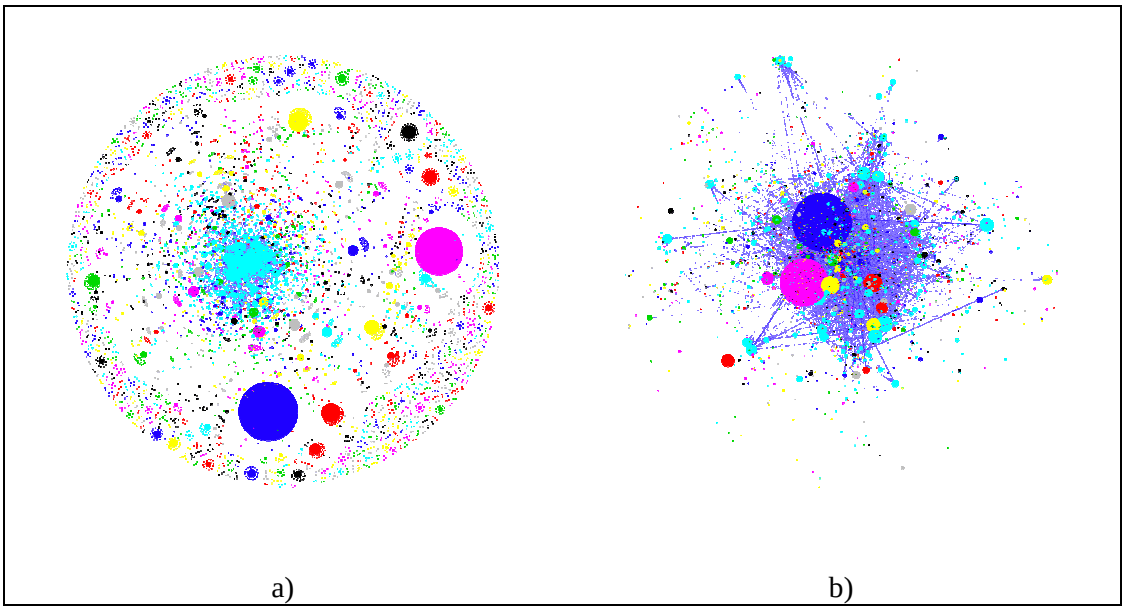


Figure 2. The network of communities, displayed a) using a Fruchterman-Reingold layout b) over geographic space, with vertices with size related to their degree.

Some examples for different communities are shown in Figure 3, where he varying cohesion and edge degree of community members are evident. Some communities, some sub-graph regions exists over a small specific geographic space, specific locales in London while other exhibit much greater spatial heterogeneity.

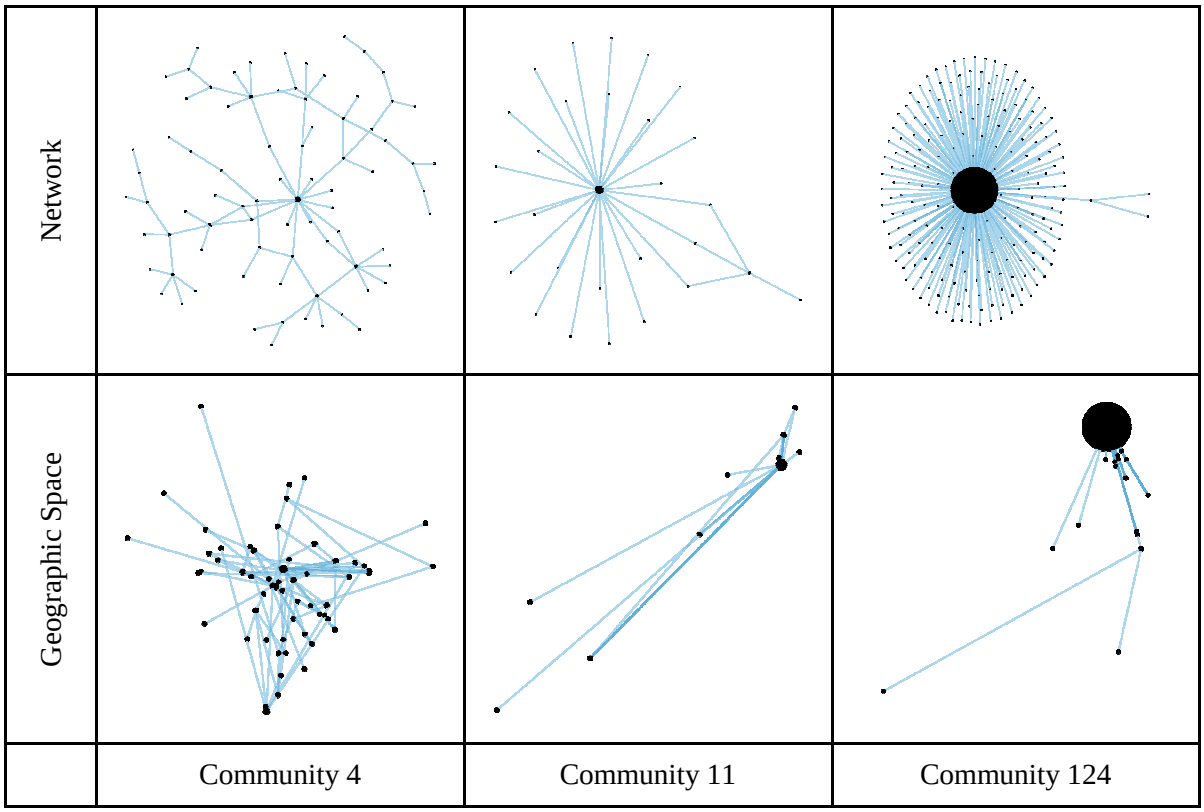
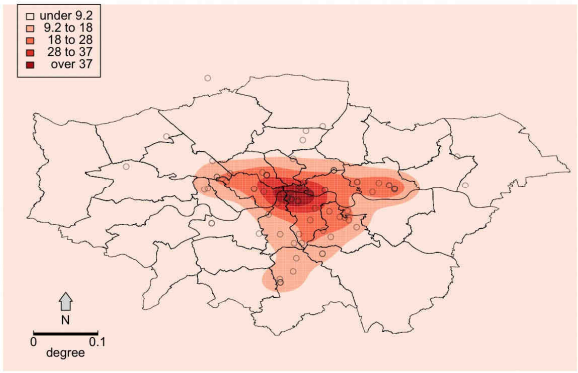
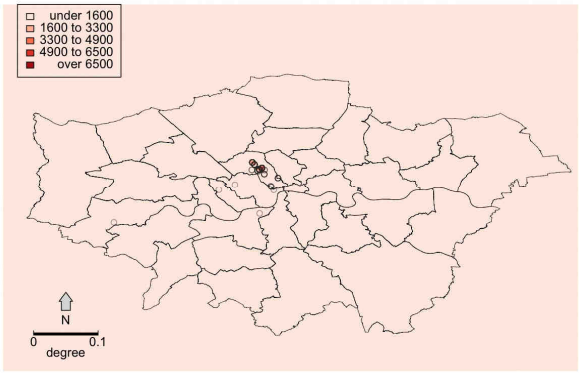


Figure 3. Examples of communities extracted using the Walktrap algorithm displayed using a Fruchterman-Reingold layout (top) and over geographic space (below), with the vertex size related to its degree.

Kernel Density Estimation (KDE) can be used to describe the spatial distribution of membership to individual communities. KDE models the relative density of the vertex locations (community members) as a surface created by summing the vertices in the sub-graph over a kernel function (a 2 dimensional distribution curve). Then for each cell, x, in a grid of discrete subdivisions of space, the relative likelihood of the occurrence of a vertex in that grid cell *f(x)* are computed. KDE's were generated for communities illustrated in Figure 4. The surfaces in Figure 4 reveal much more than simply locating the community member (vertex) locations as in Figure 4. The density surface values describe relative the intensity of the presence of the community in different locations. What they reveal are the very different spatial extents and densities of different communities. For example, Community 124 for example is relative large but very spatially concentrated in a specific area. By contrast, Community 5 is a smaller community with much greater spatial spread but with much lower density or spatial concentration of vertices.



a) Community 5



b) Community 124

Figure 4. Kernel density surfaces showing the spatial extent and spatial concentrations of for different communities.

4. Discussion

This research raises a number of discussion points:

The results the use of methods from statistical physics for partitioning networks based on the network structure, where edges indicate interactions between users. The geography of these networks can be visualised by plotting user geo-codes. Surfaces describing the relative likelihood of the occurrence of a user in any given community in any given location can be generated using kernel density estimates and future work will use text-mining of the content to communications in order to to identify connected users around specific 'themes'. Emphasising the geographic attributes of such networks provides greater insight into the structure, and possibly the activities and interests, of the communities they represent. Other areas for investigation include

- examining other partitioning algorithms (Spinglass, Leading Eigenvector, Fastgreedy and Edge Betweenness etc) ad they have been shown to be more or less suited to analysing geographic networks (Comber et a., in press);
- Exploring the use of different statistical models to determine 'surprising' arrangements of edges with the Modulaity measure – a widely used partition quality measure.
- Using methods for analysing semantics such as LDA to generate a weighted network which can be partitioned.
- Temporal / dynamic nature of social networks – need for on the fly analyses
- Exploration of heuristic search methods that explicitly deal with groups and set such as modified Grouping GAs for partitioning the network.
- Partitioning of text weighted networks

5. References

Comber A., Brunsdon, C. and Farmer, C. (in press). Community detection in spatial networks: inferring land use from a planar graph of land cover objects. Paper accepted for publication in *International Journal of Applied Earth Observation and Geoinformatio*n (January 2012)

Fortunato S (2010). Community detection in graphs. *Physics Reports*, 486(3-5): 75-174.

Girvan M and Newman MEJ (2002). Community structure in social and biological networks, *Proceedings of the National Academy of Sciences*, 99: 7821-7826.

Leicht EA and Newman MEJ (2008), Community structure in directed networks, *Physical Review Letters*, 100: 118703.

Newman MEJ (2008). The physics of networks, *Physics Today*, 61(11): 33-38.

Pons P and Latapy M (2005). Computing communities in large networks using random walks. http://arxiv.org/abs/physics/0512106v1.

Porter, MA, Onnela, J-P and Mucha, PJ, (2009). Communities in Networks. *Notices of the AMS*,

56(9): 1082-1166.

6. Biography

Lex Comber and Chris Brunsdon have research interests in spatial analysis geocomputation. Lex is now a boring old git and Chris can claim to have been to every GISRUK that ever was. Michael Batty is Emeritus Professor of Planning at the Centre for Advanced Spatial Analysis and has a long interest in all things urban and spatial. Fabian Neuhaus is a researcher at the Bartlett Centre for Advanced Spatial Analysis examining the temporal aspects of the city, focusing on routine and repetition. Andrew Hudson-Smith has an extensive research interest in digital technologies for communication. Steven Gray is a researcher at the Bartlett Centre for Advanced Spatial Analysis examining the temporal aspects of the city, focusing on routine and repetition.